S.I.E.V.E. Progress Report

Graham Clenaghan
Nick Kullman
Wayne Yang

Spring 2015

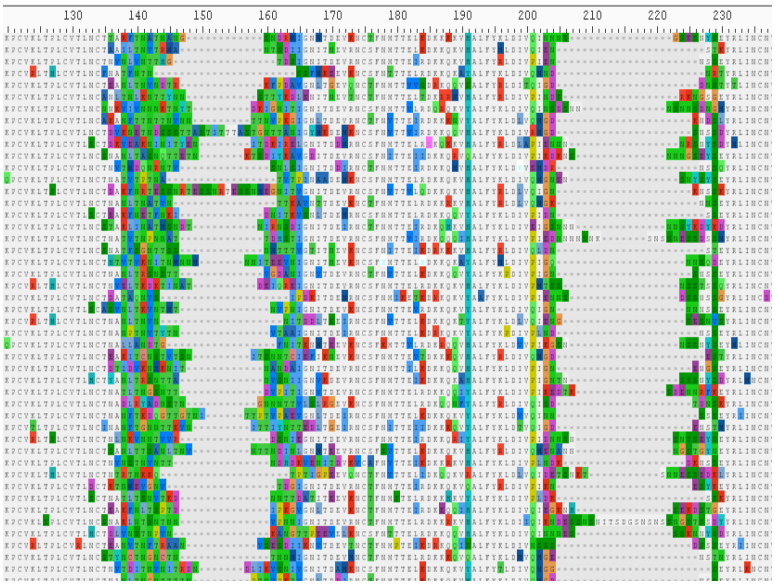
Sieve Analysis for Vaccine Trials

Sieve analysis is a statistical tool that aims to improve vaccine efficacy by providing information regarding how vaccine efficacy depends on characteristics of an exposing pathogen. Since the mid 1990s, it has been used in vaccine trials for cholera, HIV-1, hepatitis B, rotavirus, and pneumococcus. The metaphorical "sieve" in sieve analysis is the vaccine's genomic / proteomic sequence-specific immunity barrier to disease. The pathogen penetrates the vaccine's immunity barrier through "holes" in the "sieve" to cause disease [1]. Determining which characteristics of the pathogen's genomic / proteomic sequence allow it to pass through the holes will suggest antigens to include in future vaccine constructions to fight the pathogen. Determining these characteristics is the purpose of sieve analysis.
Our project aims to help vaccine researchers explore data from trials and and aid in performing sieve analysis of the results, which involves a comparison of the genomes of strains infecting patients administered a vaccine and patients administered a placebo.

While there are some existing visualizations, none of them completely fulfill this task and do not allow easy manipulation of the graphic and exploration of the data at the level we hope to accomplish. We also aim to make the tool extensible and the final product polished enough to be used by researchers in the field for use in publications in further vaccine studies.

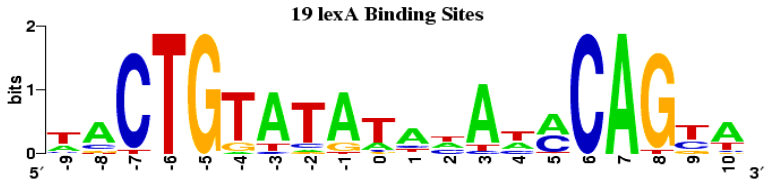
Sequence Data Visualization Tools

There are several existing tools for the visualization of genomic / proteomic sequence data. Some of these tools tend to provide a very detailed display of the alignments of sequences for a particular gene / protein from multiple patients. While this approach allows a user to see all of their data at once, it does not provide quick and easy analysis of particular sites in the sequence across patients or subsets of patients. The following example is from a software called Aliview [2].

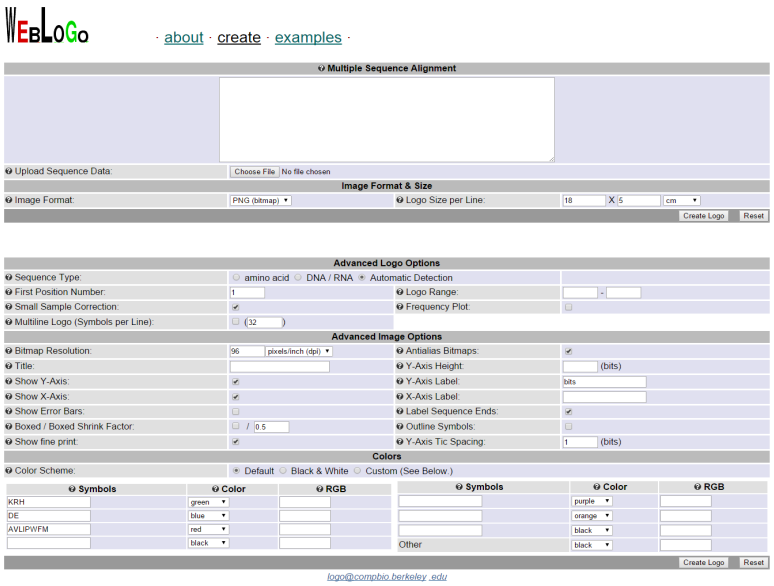


Clearly, understanding the pattern of mutations and any relationships to treatment status for a particular site in the sequence is very difficult to do in this view with any precision.

Other tools are geared towards specific analyses of sequence data. The following figure was generated using WebLogo [3], which is an online tool that can be used to parse sequence data and generate plots:



Ignoring some of the aesthetic choices, the main issue is that the primary user interface looks like this:

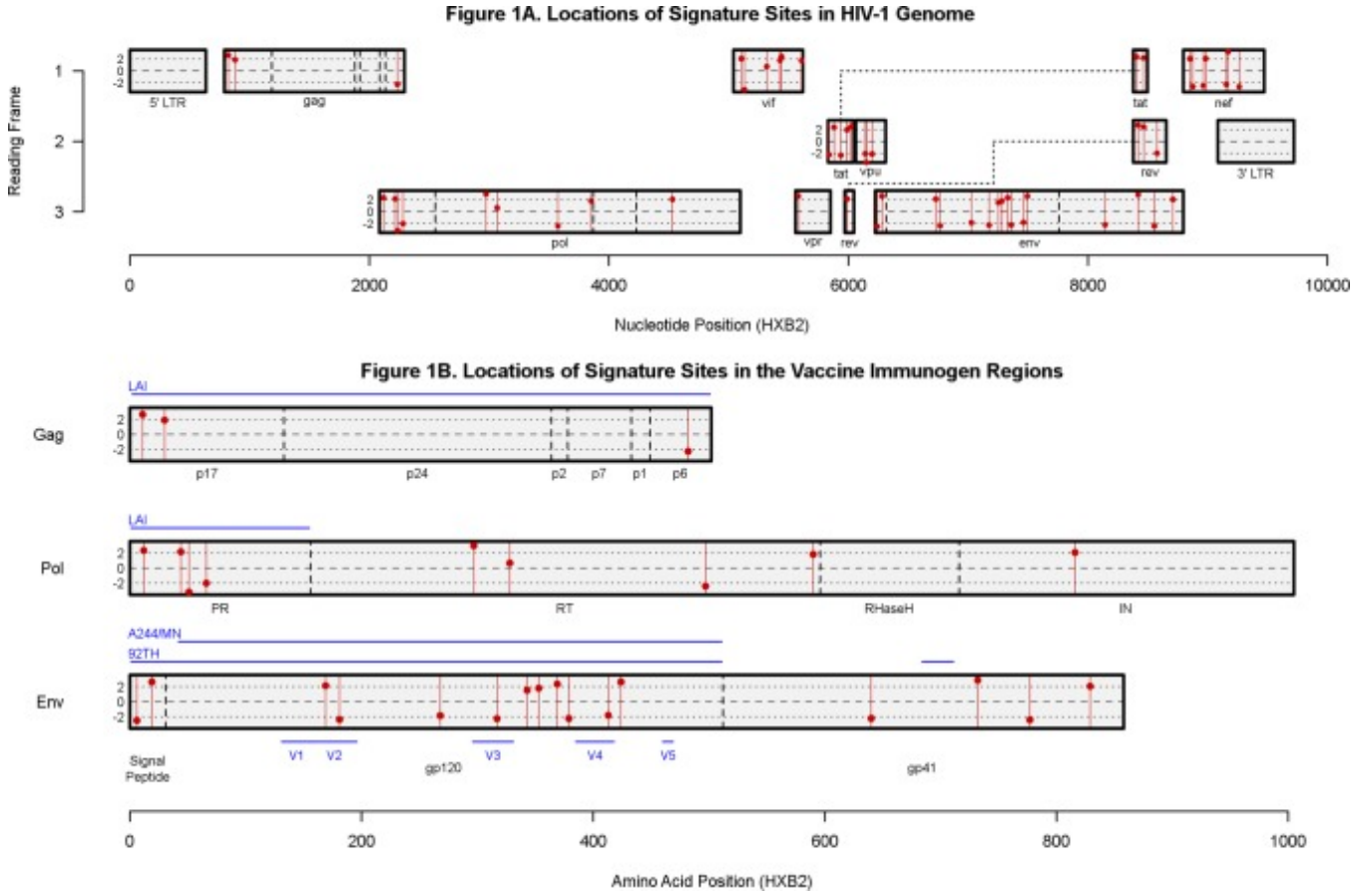


While a graphic can be automatically generated, the user has to be quite specific about what is actually plotted. As a result, if there are some sites that are known to be interesting, then a researcher can use WebLogo to construct plots. However, this system makes exploratory data analysis and actually browsing the visualization

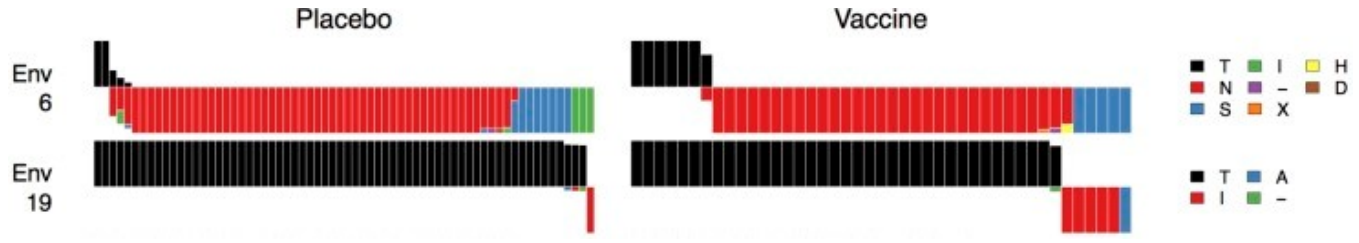
rather difficult.

Another tool that came up in our discussions with Dr. Gartland is the HIV Genome Browser, which allows users to browse reference data hosted online [4]. The browser is highly interactive and allows a user to explore the available sequences. However, this visualization does not easily allow a user to import their own data and compare across patients and treatment groups as required by sieve analysis.

The main inspiration for the project is from the paper [5], which performs sieve analysis on an HIV vaccine trial. Graphics generated in this paper include an overview of the HIV genome with statistically significant sites annotated:



and charts of individual sites showing mismatch prevalence.



These graphics were generated by a python script which requires manual coding to generate each plot. We hope to improve on this by creating an interactive tool to aid in graphic generation and aid in exploring the data visually.

Project Plan

A work-in-progress prototype is available here: `http://cse512-15s.github.io/fp-nkullman-gclenagh-wfyang/`. The utility currently supports a very basic selection mechanism of amino acid sites, which generates charts showing the prevalence of mismatches between the HIV strains found in patients and those found in the vaccine itself, and a chart showing the distribution of distances from the vaccine to the patient strains. These charts form the basic analysis tools for sieve analysis, which has the hypothesis that the strains which patients are infected with will differ between the vaccine and placebo groups at statistically significant rates.

The remaining tasks are roughly:

- Finalize the site selection mechanism / interface. Primary member: Nick.
- Color palette selection menu. Primary member: Graham.
- Incorporate p-values / entropy calculations. Primary member: Wayne.
- Enable exporting of graphics.
- General polish / optimizations.

References

[1] Peter Gilbert, Steve Self, Malla Rao, Abdollah Naficy, and John Clemens. Sieve analysis: methods for assessing from vaccine trial data how vaccine efficacy varies with genotypic and phenotypic pathogen variation. *Journal of clinical epidemiology*, 54(1):68–85, 2001.

[2] A. Larsson. Aliview: A fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics*, 30, 2014.

[3] J.M. Chandonia G. Crooks, G. Hon and S. Brenner. Weblogo: A sequence logo generator. *Genome Research*, 14, 2004.

[4] Hiv genome browser. http://www.hiv.lanl.gov/.

[5] Paul T Edlefsen, Morgane Rolland, Tomer Hertz, Sodsai Tovanabutra, Andrew J Gartland, Allan C deCamp, Craig A Magaret, Hasan Ahmed, Raphael Gottardo, Michal Juraska, et al. Comprehensive sieve analysis of breakthrough hiv-1 sequences in the rv144 vaccine efficacy trial. *AIDS research and human retroviruses*, 30(S1):A25–A26, 2014.